40386

National Library of Canada | Bibliothèque nationale du Canada

CANADIAN THESES ON MICROFICHE | THÈSES CANADIENNES SUR MICROFICHE

NAME OF AUTHOR/*NOM DE L'AUTEUR* Daniel Kam-Kui Chow

TITLE OF THESIS/*TITRE DE LA THÈSE* On the Construction of Feedback Queues

UNIVERSITY/*UNIVERSITÉ* U. of Alberta

DEGREE FOR WHICH THESIS WAS PRESENTED/ *GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE* M.Sc.

YEAR THIS DEGREE CONFERRED/*ANNÉE D'OBTENTION DE CE GRADE* 1979

NAME OF SUPERVISOR/*NOM DU DIRECTEUR DE THÈSE* C.T. Yu

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

DATED/*DATÉ* January 2, 1979 SIGNED/*SIGNÉ* K. K. Chow

PERMANENT ADDRESS/*RÉSIDENCE FIXÉ* 26 A Tai Ping Shan Street
5/F Sheung Wan
Hong Kong

NL-91 (3-74)

National Library of Canada

Cataloguing Branch
Canadian Theses Division

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada

Direction du catalogage
Division des thèses canadiennes

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

THE UNIVERSITY OF ALBERTA

ON THE CONSTRUCTION OF FEEDBACK QUERIES

BY

DANIEL KAM-KUI CHOW

(C)

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTING SCIENCE

EDMONTON, ALBERTA

SPRING 1979

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, the thesis entitled ON THE CONSTRUCTION OF FEEDBACK QUERIES submitted by Daniel Kam-Kui Chow in partial fulfilment of the requirements for the degree of Master of Science.

.....
(Supervisor)

.....

.....

Date December 8, 1978

Abstract

The purpose of this thesis is to construct optimal feedback queries in information retrieval systems. An optimal retrieval rule is derived using the Neyman-Pearson decision rule. Three probabilistic models and the optimal queries to be used in the models are presented. Parameters which are required to construct these queries are estimated, based on relevance information from the user about the retrieved documents. Finally, the effect of deleting a term from the optimal query in one of the three models on retrieval performance is analysed.

ACKNOWLEDGEMENTS

I wish to thank Dr. C.T. Yu, my supervisor, for his advice and guidance throughout the preparation of this thesis.

Thanks are due to Dr. T.A. Marsland, Dr. C.S. Hoo and my fellow graduate Miss A. Brindle for their criticisms and valuable suggestions. I am also indebted to Miss Grace Ting for her invaluable assistance in typing this thesis.

The financial assistance received from the Department of Computing Science and the National Research Council, in the form of teaching and research assistantship, is appreciated.

TABLE OF CONTENTS

	Page
Chapter 1: Introduction	1
Chapter 2: The System, Performance Measure and Optimal Retrieval Rule	4
2.1 The System	4
2.2 Performance Measure - Recall and Precision ..	5
2.3 Optimal Retrieval Rule	6
Chapter 3: The Models and Optimal Queries	9
3.1 Model 1	9
3.2 Model 2	11
3.3 Model 3	12
Chapter 4: Estimation of Parameters in the Three models	15
4.1 Introduction to Parameter Estimations	15
4.2 Estimation of Parameters in Model 1	17
4.3 Estimation of Parameters in Model 2	21
4.4 Estimation of Parameters in Model 3	24
4.5 Estimation of the Number of Relevant Documents	29

Chapter 5: Measurement of the Importance of Index Term with respect to Retrieval Performance in Model 1 30

5.1 Performance Measure 30

5.2 Q*-{j} is not Better than Q*-{i} at all Recall Levels 33

5.3 Which is Better, Q*-{j} or Q*-{i} ? 36

Chapter.6: Conclusion 47

References 50

Appendix 1 53

Chapter 1

Introduction

The purpose of constructing feedback queries [14] is to give the user more relevant documents. Documents and queries in information retrieval systems are usually represented by n-dimensional vectors whose components are the keywords (keywords, index terms and terms will be used interchangeably). An example is the SMART System [15]. In response to a user query Q, the system retrieves documents that are "close" to the query . A simple way to measure the "closeness" or "retrieval status value" of a document with respect to a query is the number of keywords or terms in common between them. The terms may be weighted with respect to their importance in the vector. So we can also consider a document or a query as a set of weighted terms.

Since what the user wants can seldom be fully expressed by a set of keywords (i.e. his query) and the notion of closeness may not be exactly represented by the measure indicated above, it is usual that only some of the retrieved documents are found to be of interest to him. If the user is not satisfied with the retrieved documents, he may request the system to reformulate the query by indicating which of the retrieved documents are relevant to him. The

system then attempts to make use of the feedback information to obtain better retrieval performance.

Rocchio and Salton [14] suggest a practical method for modifying queries to achieve better performance. The new query is :

$$Q' = Q + a * \sum_{x \in Q(R)} \underline{x} - b * \sum_{x \in Q(I)} \underline{x}'$$

where $a \geq 0$, $b \geq 0$ are parameters to be determined and Q(R) and Q(I) are respectively the set of relevant documents and irrelevant documents retrieved by the initial query, Q.

Yu, Luk and Cheung [18] analyse necessary and sufficient conditions under which the parameters a and b make Q' a better query than Q. The exact values of a and b to produce optimal results are not located. Their experimental results indicate that choosing $a=1/|Q(R)|$ and $b=1/|Q(I)|$ works well in practice.

The focus of previous research [13,17] is mainly on the optimality of queries of model 1 presented in chapter 3 of this thesis. Robertson and Sparck Jones [13] report an optimal query by the Bayes' theorem. Van Rijsbergen [17] uses Bayes decision rule to derive similar results of [13] and he further extends the model to a special case of term dependence. Kraft and Bookstein [9] use the Neyman-Pearson [10] lemma to determine the optimal range of retrieval status value.

This thesis analyses optimal queries in information retrieval systems. In chapter 2, an optimal retrieval rule is derived using the Neyman-Pearson Decision Rule [6]. In chapter 3, this optimal retrieval rule is applied to three rather common models in information retrieval. In each case, an optimal query is derived. In chapter 4, the necessary parameters required to construct the optimal queries in the three models are estimated, based on relevance information supplied by the user. In reference [13,17] an attempt is made to estimate the parameters in model 1 based on relevance information of random documents. However, the estimations in chapter 4 make use of retrieved documents. In chapter 5, the effect of deleting a term from the optimal query in model 1 is analysed. A scheme to rank the usefulness of the terms in retrieval is proposed. It is hoped that the same approach can be generalized to the case of deleting more index terms and to other models. The deletion of terms from feedback queries or queries containing too many index terms is necessary in order to speed up the process of retrieval.

Chapter 2

The System, Performance Measure and Optimal Retrieval Rule

2.1 The System

Documents and queries are represented by n-dimensional vectors whose components are associated with the index terms of the document space. It is assumed that the higher the value of the i^{th} component, the more important is the i^{th} term to the document. The value of a term in a vector is known as its weight.

The system retrieves documents that are 'close' to the query. The 'closeness' or 'retrieval status value' of a document vector $\underline{x}$ with respect to the query vector Q is measured by the system by means of a real-valued function f. A query Q retrieves document $\underline{x}$ if and only if $f(Q,\underline{x}) > K$ where K is a threshold value. In other words, if $\underline{x}$ satisfies $f(Q,\underline{x}) > K$ then $\underline{x}$ is assumed by the system to be close to the query Q.

Heine [8] has discussed several forms of f. Here a simple matching function wil be used :

$$f(Q,\underline{x}) = \sum_{i=1}^{n} q_i * x_i \qquad (2.1.1)$$

$= Q.\underline{x}$ (dot product)

$= Q * \underline{x}^T$ (vector multiplication)

where $Q = (q_1, q_2, \ldots, q_n)$

$\underline{x} = (x_1, x_2, \ldots, x_n)$

and $\underline{x}^T$ is the transpose of $\underline{x}$.

2.2 Performance Measure - recall and precision

Whether a document is relevant or not relevant to a query is entirely dependent on the user. An ideal system retrieves all relevant documents and no irrelevant documents. However, retrieval by means of a closeness function rarely yields the desired result. Thus the objective is to retrieve as few irrelevant documents as possible while retrieving a certain number of relevant documents. This may be phrased in terms of the two most common retrieval performance measures in information retrieval, recall and precision, defined as follows:

Recall = Probability that a document is retrieved given that the document is relevant,

Precision = Probability that a document is relevant given that the document is retrieved.

It is clear that the above objective is equivalent to obtaining the highest precision at any given level of

recall. We now present a retrieval strategy that achieves this aim. This strategy will be shown to be equivalent to minimizing type 2 (beta) error for fixed type 1 (alpha) error in statistical decision theory.

2.3 Optimal Retrieval Rule

Let

P(A) be the probability that event A occurs,

P(A|B) be the conditional probability of occurrence of the event A, given the event B,

P(A,B) be the probability that events A and B co-occur,

DR be the set of documents retrieved by Q,

R be the set of documents relevant to the query Q,

I be the set of documents irrelevant to the query Q,

D be the set of documents in the document space,

|Y| be the number of elements in the set Y,

C_1 be the event that a document is relevant to Q,

C_2 be the event that a document is irrelevant to Q,

dr be the event that a document is retrieved by Q,

θ be the probability that a randomly chosen document is relevant to the query Q,

i.e. $\theta = |R|/|D|$ or $P(C_1)$,

$\phi(\underline{x})$ be the probability that document $\underline{x}$ is relevant to the query Q,

i.e. $P(C_1|\underline{x})$.

Let

alpha = $P(\tilde{}dr|C_1)$

and beta $= P(dr|C_2)$.

Then

recall = 1 - alpha

precision $= P(C_1|dr)$

$= P(dr|C_1) * P(C_1) / P(dr)$

$= P(dr|C_1) * P(C_1) /$

$(P(dr|C_1)*P(C_1) + P(dr|C_2)*P(C_2))$

$= (1-alpha)*\Theta / ((1-alpha)*\Theta+beta*(1-\Theta))$

where $\Theta = P(C_1)$.

Since Θ is independent of any retrieval strategy, it is easy to see :

at fixed alpha, beta is minimized if and only if

at fixed recall level, precision is maximized.

Therefore, the objective of maximizing precision at any given level of recall can be met by the Neyman-Pearson Decision Rule [6] and the retrieval rule that achieves the objective is :

$$\text{retrieve } \underline{x} \text{ if and only if } \frac{P(\underline{x}|C_1)}{P(\underline{x}|C_2)} > K \qquad (2.3.1)$$

The following shows that the retrieval rule is optimal, in the sense that it can rank documents in order of their probability of relevance.

The relation between the probability of relevance, $\phi(\underline{x})$, of document $\underline{x}$, and the ratio $P(\underline{x}|C_1)/P(\underline{x}|C_2)$ is :

$$\phi(\underline{x}) = P(C_1|\underline{x})$$

$$= P(\underline{x}|C_1) * P(C_1) / P(\underline{x})$$

$$= P(\underline{x}|C_1) * P(C_1) / (P(\underline{x}|C_1)*P(C_1) + P(\underline{x}|C_2)*P(C_2))$$

$$= P(\underline{x}|C_1) * \Theta / (P(\underline{x}|C_1)*\Theta + P(\underline{x}|C_2)*(1-\Theta))$$

$$= P(\underline{x}|C_1) / P(\underline{x}|C_2) * \Theta / (P(\underline{x}|C_1)/P(\underline{x}|C_2)*\Theta+(1-\Theta))$$

We immediately have : (for any documents $\underline{x}_1$, $\underline{x}_2$)

$\phi(\underline{x}_1) > \phi(\underline{x}_2)$ if and only if

$P(\underline{x}_1|C_1)/P(\underline{x}_1|C_2) > P(\underline{x}_2|C_1)/P(\underline{x}_2|C_2)$

Hence, $P(\underline{x}|C_1)/P(\underline{x}|C_2)$ also ranks documents in order of their probability of relevance. As a consequence, retrieving documents in decreasing order of $P(\underline{x}|C_1)/P(\underline{x}|C_2)$ yields the highest expected number of relevant documents for any given number of documents retrieved.

Next chapter, we shall construct optimal feedback queries under three different distributions of term weights on the documents, where an optimal feedback query Q^*,

$$Q^* = (w_1, w_2, \ldots, w_n)$$

is a query satisfying

$$\frac{P(\underline{x}|C_1)}{P(\underline{x}|C_2)} \iff f(Q^*,\underline{x}_1) > f(Q^*,\underline{x}_2)$$ for any documents $\underline{x}_1$, $\underline{x}_2$

i.e. a query Q^* that can rank documents in order of their probablity of relevance.

Chapter 3

The Models and Optimal Queries

This chapter presents three commonly used models in information retrieval. The optimal queries in the models are derived, making use of the optimal retrieval rule.

Let T_i, $1 \leq i \leq n$, be a random variable associated with the weights of the i^{th} term. The models are characterized by the distributions assumed on $\underline{T} = (T_1,\ldots,T_n)$.

3.1 Model 1

Each document is a binary vector, i.e. its i^{th} component is 1 or 0, depending respectively on the presence or the absence of the i^{th} term in the document, $1 \leq i \leq n$. Furthermore, when restricted to the relevant documents of Q, the T_i's are mutually independent; when restricted to the irrelevant documents of Q, the T_i's are also mutually independent.

This model has been used in the analysis of various information retrieval processes [13,19,21]. Optimal queries derivable from this model are found to be "first order" or linear approximations to those obtainable in a binary model

which incorporates the dependence of terms [Appendix 1].

Let $p_i = P(T_i=1|C_1)$

$r_i = P(T_i=1|C_2)$

and $\underline{x} = (x_1, \ldots, x_n)$ be a document

Then, by the independence of the T_i's and binary nature of the term weights,

$$\frac{P(\underline{T}=\underline{x}|C_1)}{P(\underline{T}=\underline{x}|C_2)} = \frac{\prod_{i=1}^{n} p_i^{x_i} (1-p_i)^{1-x_i}}{\prod_{i=1}^{n} r_i^{x_i} (1-r_i)^{1-x_i}}$$

$$= \prod_{i=1}^{n} \left(\frac{p_i/(1-p_i)}{r_i/(1-r_i)} \right)^{x_i} * \frac{1-p_i}{1-r_i}$$

$$\Rightarrow \log \frac{P(\underline{T}=\underline{x}|C_1)}{P(\underline{T}=\underline{x}|C_2)} = \sum_{i=1}^{n} \log\left(\frac{p_i/(1-p_i)}{r_i/(1-r_i)}\right) * x_i + \sum_{i=1}^{n} \log \frac{1-p_i}{1-r_i}$$

Let

$$w_i = \log \frac{p_i/(1-p_i)}{r_i/(1-r_i)}$$

$$\log \frac{P(\underline{T}=\underline{x}|C_1)}{P(\underline{T}=\underline{x}|C_2)} = \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \log \frac{1-p_i}{1-r_i}$$

Let $Q^* = (w_1, w_2, \ldots, w_n)$

Then $f(Q^*, \underline{x}) = \sum_{i=1}^{n} w_i x_i$

$$= \log \frac{P(\underline{T}=\underline{x}|C_1)}{P(\underline{T}=\underline{x}|C_2)} - \sum_{i=1}^{n} \log \frac{1-p_i}{1-r_i}$$

It is clear that for any two documents $\underline{x}_1$ and $\underline{x}_2$

$f(Q^*,\underline{x}_1) > f(Q^*,\underline{x}_2)$ if and only if

$$\log \frac{P(\underline{T}=\underline{x}_2|C_1)}{P(\underline{T}=\underline{x}_1|C_2)} > \log \frac{P(\underline{T}=\underline{x}_2|C_1)}{P(\underline{T}=\underline{x}_2|C_2)}$$

Since log is a monotonic increasing function of its arguments, the ranking of documents by Q* is equivalent to that by the optimal retrieval rule. In other words, with respect to model 1, Q* is an optimal query.

Yu and Salton [20] suggest ranking terms in decending order of $[p_i/(1-p_i)]/[r_i/(1-r_i)]$. Roberson and Sparck Jones [13] show that taking the logarithm of that expression gives an optimal result, although the derivation is different from that given here. Van Rijisbergen [17] uses the Bayes decision rule to derive the same result.

3.2 Model 2

In this model, the frequency of occurrence of any term in the relevant documents follows a Poisson distribution [3,7]; its distribution in the irrelevant documents is also Poisson with a different parameter. Furthermore, the frequencies of occurrences of the terms in the set of relevant documents (and in the set of irrelevant documents) are independent. This can be considered as a slight modification of the linked-two-Poisson model [2].

More precisely, when restricted to R, each T_i takes on

a Poisson distribution with parameters u_i and all the T_i's are mutually independent. When restricted to I, each T_i is Poisson distributed with parameter v_i and the T_i's are again mutually independent. It follows that $E[T_i|C_1] = u_i$ and $E[T_i|C_2] = v_i$.

By the independence assumption,

$$\frac{P(\underline{T}=\underline{x}|C_1)}{P(\underline{T}=\underline{x}|C_2)} = \frac{\prod_{i=1}^{n} P(T_i=x_i|C_1)}{\prod_{i=1}^{n} P(T_i=x_i|C_2)}$$

$$= \frac{\prod_{i=1}^{n} (u_i^{x_i} * e^{-u_i}/x_i!)}{\prod_{i=1}^{n} (v_i^{x_i} * e^{-v_i}/x_i!)}$$

$$= \exp\left[\sum_{i=1}^{n} (v_i - u_i) \right] * \prod_{i=1}^{n} (u_i/v_i)^{x_i}$$

Let $w_i = \log (u_i/v_i)$

$$\log \frac{P(\underline{T}=\underline{x}|C_1)}{P(\underline{T}=\underline{x}|C_2)} = \sum_{i=1}^{n} (v_i - u_i) + \sum_{i=1}^{n} w_i x_i$$

$Q^* = (w_1, w_2, \ldots, w_n)$.

It is clear that Q^* is an optimal query with respect to model 2.

3.3 Model 3

Here the joint distribution of the weights of the terms in the relevant documents is multivariate normal [11]; in

the irrelevant documents, it is also multivariate normal.

Under this assumption, $f(Q,\underline{x})$ is normally distributed [11]. This model is consistent with the Swets model [16], though the assumption here is somewhat stronger.

Recently, there have been some doubts about the validity of assuming normal distribution for term weights over documents. However, if the correlations of terms are removed by some procedures such as factor analysis [1,4,12], then it is very likely that new terms are normally and independently distributed.

A more precise characterization of the model is as follows. When restricted to the relevant documents, $\underline{T}$ has an n-dimensional normal distribution with mean vector $\underline{u} = (u_1, \ldots, u_n)$ and covariance matrix $\underline{S}_1$; when restricted to the irrelevant documents, it has an n-dimensional normal distribution with mean vector $\underline{v} = (v_1, \ldots, v_n)$ and covariance matrix $\underline{S}_2$.

By the hypothesis, $u_i = E[T_i|C_1]$ and $\underline{S}_1(i,j)$ is the covariance of T_i and T_j when restricted to the relevant documents. The optimal retrieval rule for the continuous case is :

retrieve $\underline{x}$ if and only if $\dfrac{p(\underline{x}|C_1)}{p(\underline{x}|C_2)} > K$

where

$p(\underline{x}|C_1)$ and $p(\underline{x}|C_2)$ are the probability density

functions of $\underline{T}$ when restricted to the relevant documents and irrelevant documents respectively.

Let PI = 3.1416

$|A|$ be the determinant of matrix A

A^{-1} be the inverse of matrix A

and $\underline{x}^T$ be the transpose of vector $\underline{x}$.

Then

$$\frac{p(\underline{x}|C_1)}{p(\underline{x}|C_2)} = \frac{(2*PI)^{-n/2}|\underline{S}_1|^{-1} \exp\{-1/2(\underline{x}-\underline{u})^T\underline{S}_1^{-1}(\underline{x}-\underline{u})\}}{(2*PI)^{-n/2}|\underline{S}_2|^{-1} \exp\{-1/2(\underline{x}-\underline{v})^T\underline{S}_2^{-1}(\underline{x}-\underline{v})\}} .$$

It follows that ranking document $\underline{x}$ by

$$(\underline{x}-\underline{u})^T\underline{S}_1^{-1}(\underline{x}-\underline{u}) - (\underline{x}-\underline{v})^T\underline{S}_2^{-1}(\underline{x}-\underline{v}) \qquad (3.3.1)$$

is equivalent to ranking $\underline{x}$ by $p(\underline{x}|C_1)/p(\underline{x}|C_2)$.

There are two important subcases of (3.3.1).

(i) when $\underline{S}_1 = \underline{S}_2 = \underline{S}$

the ranking procedure can be reduced to a linear form :

$$\underline{x}^T\underline{S}^{-1}(\underline{u}-\underline{v}) \qquad (3.3.2)$$

which is the well known Fisher linear discriminant [5] for general 2-way classification problems.

(ii) when $\underline{S}_1 = \underline{S}_2 = \underline{S}$ and all index terms are independent, (3.3.2) can be further reduced to

$$\sum_{i=1}^{n} w_i x_i$$

where

$$w_i = (f_i - v_i)/s_i^2$$

$$f_i = E[T_i]$$

$$s_i^2 = \mathrm{Var}[T_i].$$

and $Q^* = (w_1, w_2, \ldots, w_n)$ is the optimal query under the assumptions.

Chapter 4

Estimation of Parameters in the 3 Models

4.1 Introduction to Parameter Estimations

In a realistic retrieval environment, the user specifies the content of what he intends to retrieve by a set of keywords. These keywords may be weighted by the user and/or the system. However the user may not be satisfied with the performance of the initial query Q. the keywords specified. The system will be required to modify the query Q after the user identifies the relevant documents in the retrieved set of documents.

In this chapter, it is assummed that

$Q = (q_1, q_2, \ldots, q_r, 0, \ldots, 0)$

where each q_i is a positive integer for $i \leq r$.

The optimal queries of the three models have been discussed and derived in chapter 3. The following parameters can be estimated, and the estimation process will be presented in this chapter.

p_i, r_i in model 1

u_i, v_i in model 2

$\underline{u}$, $\underline{v}$, $\underline{S}_1$, $\underline{S}_2$ in model 3

These parameters are needed to construct the optimal queries in the three models. Furthermore, when the above parameters have been estimated, the number of relevant documents in the collection can also be estimated.

The estimation process for the parameters in R (the set of relevant documents) is the same as that for the parameters in I (the set of irrelevant documents). Hence random variables defined in this chapter are restricted to R unless explicitly stated. Thus T_i will be taken to be $T_i|C_1$.

Let

$$u_i' = E[T_i|dr] \text{ and } p_i' = P(T_i=1|dr).$$

After the user identifies the relevant documents in the retrieved set of documents, both u_i' and p_i' can easily be estimated. Clearly, when T_i takes on binary values, $u_i' = p_i'$.

The following proposition shows that the parameters u_i and p_i may be estimated as u_i' and p_i' in models 1 and 2 for those terms absent in the original query Q.

Proposition 4.1.1

If the T_j's, $1 \leq j \leq n$, are independent,

then $u_i' = u_i$ for $i > r$.

Proof :

By definition,

$$u_i = \sum_t t * P(T_i=t) \text{ and}$$

$$u_i' = \sum_{t} t * P(T_i=t|dr).$$

It is sufficient to show

$P(T_i=t) = P(T_i=t|dr)$ or equivalently

$P(dr) = P(dr|T_i=t)$ for any t.

$$P(dr|T_i=t) = \sum_{c=K+1}^{\infty} P\left(\sum_{j=1}^{r} q_j T_j = c \,|\, T_i=t\right)$$

$$= \sum_{c=K+1}^{\infty} P\left(\sum_{j=1}^{r} q_j T_j = c\right)$$

$$= P(dr) .$$

4.2 Estimation of Parameters in Model 1

The parameters to be estimated are the p_i's. The next proposition shows that if the c^{th} term is independent of "sufficiently many" terms of the original query, then p_c can be estimated. This is a weaker independence assumption than the one in Model 1.

Proposition 4.2.1

Let $Q = (q_1, q_2, \ldots, q_r, 0, \ldots, 0)$ and consider Q as a set of keywords, $Q = \{1,2, \ldots ,r\}$.

Let S_c, $S_c \subseteq Q$, be a set of terms of size m.

If T_c and $\sum_{i \in S_c} q_i T_i$ are independent

and $\sum_{i \in S_c} q_i > K$, then

$$p_c = \frac{p_c' * |R \cap DR| - \overline{\sum_{j \in Q-S_c}} E_j}{A_c} \qquad (4.2.1)$$

where E_j , $j \in Q-S_c$, is the expected number of relevant documents containing terms c and j and satisfying

$$K \geq \overline{\sum_{i \in S_c}} q_i T_i + \overline{\sum_{s \in Q-S_c, s>j}} q_s T_s > K-q_j$$

and A_c is the expected number of relevant documents satisfying $\overline{\sum_{i \in S_c}} q_i T_i > K$.

Proof :

Without loss of generality, let $S_c = \{r-m+1, \ldots ,r\}$ and $Q-S_c = \{1, \ldots ,r-m\}$. The probability that a document is retrieved by Q can be expressed by sum of probabilities as follows.

By definition,

$$P(dr) = P\left(\overline{\sum_{i=1}^{r}} q_i T_i > K\right)$$

$$= P\left(T_1=1, \overline{\sum_{i=2}^{r}} q_i T_i > K-q_1\right) + P\left(T_1=0, \overline{\sum_{i=2}^{r}} q_i T_i > K\right)$$

$$= P\left(T_1=1, K \geq \overline{\sum_{i=2}^{r}} q_i T_i > K-q_1\right) +$$

$$P\left(T_1=1, \overline{\sum_{i=2}^{r}} q_i T_i > K\right) + P\left(T_1=0, \overline{\sum_{i=2}^{r}} q_i T_i > K\right)$$

$$= P(T_1=1,\ K \geq \sum_{i=2}^{r} q_i T_i > K-q_1) + P(\sum_{i=2}^{r} q_i T_i > K)$$

expand $P(\sum_{i=2}^{r} q_i T_i > K)$ repeatedly,

$$= \sum_{j=1}^{r-m} P(T_j=1,\ K \geq \sum_{i=j+1}^{r} q_i T_i > K-q_j) + P(\sum_{i=r-m+1}^{r} q_i T_i > K) \qquad (4.2.2)$$

Then by the independence of T_c and $\sum_{i=r-m+1}^{r} q_i T_i$,

$P(dr|T_c=1)$

$$= \sum_{j=1}^{r-m} P(T_j=1,\ K \geq \sum_{i=j+1}^{r} q_i T_i > K-q_j \mid T_c=1) + P(\sum_{i=r-m+1}^{r} q_i T_i > K)$$

By definition,

$$p_c' = P(T_c=1|dr) = \frac{P(dr|T_c=1) * P(T_c=1)}{P(dr)}$$

$$= \frac{\sum_{j=1}^{r-m} P(T_j=1,\ K \geq \sum_{i=j+1}^{r} q_i T_i > K-q_j,\ T_c=1) + P(\sum_{i=r-m+1}^{r} q_i T_i > K) * p_c}{P(dr)}$$

Multiplying both the numerator and the denominator by $|R|$ and solving for p_c, the desired result follows.

It is clear that documents containing term j, $j \leq r$,

and satisfying

$$K \geq \overline{\sum_{i \in S_c}} q_i T_i + \overline{\sum_{s \in Q-S_c, s>j}} q_s T_s > K - q_j$$

are retrieved by Q, as are documents satisfying $\overline{\sum_{i \in S_c}} q_i T_i > K$.

However A_c has to be greater than zero in order to estimate p_c. In a normal retrieval situation, K is much less than $\overline{\sum_{i \in Q}} q_i$. So it should be easy to choose $S_c \subseteq Q$ such that $\overline{\sum_{i \in S_c}} q_i > K$. Thus the E_j's, A_c and hence p_c can be estimated.

In the special case where all terms are independent (which is the assumption of model 1), set $S_c = Q$ for $c > r$ and $S_c = Q-\{c\}$ for $1 \leq c \leq r$.

In order to study the relation between p_c and p_c', let E_j', $j \in Q-S_c$, be the expected number of relevant documents containing term j and satisfying

$$K \geq \overline{\sum_{i \in S_c}} q_i T_i + \overline{\sum_{s \in Q-S_c, s>j}} q_s T_s > K - q_j.$$

Then by equation (4.2.2)

$$A_c = |R \cap DR| - \overline{\sum_{j \in Q-S_c}} E_j',$$

equation (4.2.1) can be written in the form :

$$p_c = \frac{p_c'*|R\cap DR| - \overline{\sum_{j\in Q-S_c}} E_j}{|R\cap DR| - \overline{\sum_{j\in Q-S_c}} E_j'}$$

After rearranging,

$$(p_c'-p_c)*|R\cap DR| = \overline{\sum_{j\in Q-S_c}} E_j - p_c*\overline{\sum_{j\in Q-S_c}} E_j'$$

$$= \overline{\sum_{j\in Q-S_c}} (E_j - p_c*E_j') \qquad (4.2.3)$$

Consider the case when all terms are independent. $E_j = p_c E_j'$ for $j \neq c$. Therefore, when $c \notin Q$, the right hand side of equation (4.2.3) is zero, which is exactly the result of Proposition 4.1.1. On the other hand, when $c \in Q$, S_c is chosen to be $Q-\{c\}$, so the right hand side of equation (4.2.3) becomes $E_c - p_c E_c$, which is non-negative. Hence $p_c \geq p_c'$. In general, the sufficient condition for $p_c' \geq p_c$ is $E_j - p_c E_j' \geq 0$, which is true when T_c and E_j' co-occur frequently.

4.3 Estimation of Parameters in Model 2

The parameters to be estimated are the u_i's. By Proposition 4.1.1, $u_i = u_i'$, $i > r$. Thus it is sufficient to estimate u_i, $1 \leq i \leq r$.

Let $Y = q_1T_1 + q_2T_2 + \ldots + q_rT_r$

$u = E(Y)$

and $u_{ic} = E(T_i|Y=c)$

We futher assume that $q_i=1$, $i\leq i\leq r$.

The estimation process consists of expressing u in terms of other estimatable parameters by means of Proposition 4.3.1. Then u_i , $1 \leq i \leq r$, are expressed in terms of u.

Proposition 4.3.1

If Y is a Poisson random variable, then u can be estimated by setting c = K+2, K+3, etc. in the following equation :

$$u = \frac{E(Y|Y \geq c) * E_1}{E_2}$$

where

E_1 and E_2 are the expected number of relevant documents with similarity greater than or equal to c and (c-1) respectively.

Proof

After some manipulation,

$$P(Y=y|Y \geq c) = \frac{\frac{u^y e^{-u}}{y!}}{\sum_{j=c}^{\infty} \frac{u^j e^{-u}}{j!}}, \qquad \text{if } y \geq c$$

and

$$E(Y|Y \geq c) = u * \left[\frac{\left(\sum_{j=c-1}^{\infty} \frac{u^j e^{-u}}{j!} \right) * |R|}{\left(\sum_{j=c}^{\infty} \frac{u^j e^{-u}}{j!} \right) * |R|} \right]$$

Solving for u, the desired result follows.

Without loss of generality, u_1 is estimated by the following proposition.

Proposition 4.3.2

Let $Y = T_1 + \ldots + T_r$

$Z = T_2 + \ldots + T_r$

$u_{1c} = E[T_1 | Y=c]$.

If Z and T_1 are independent Poisson random variables, then u_1 can be estimated by

$$u * \frac{u_{1c}}{c}, \quad \text{for } c = K+1, K+2, \text{ etc.}$$

Proof

Since T_1 and Z are independent Poisson random variables with parameters u_1 and $u-u_1$ respectively,

$$P(T_1=t, Z=z) = \frac{u_1^t e^{-u_1}}{t!} * \frac{(u-u_1)^z e^{-(u-u_1)}}{z!}$$

=>

$$P(T_1=t, Y=y) = \frac{u_1^t e^{-u_1}}{t!} * \frac{(u-u_1)^{y-t} e^{-(u-u_1)}}{(y-t)!}$$

$$P(T_1=t|Y=y) = \frac{P(T_1=t, Y=y)}{P(Y=y)}$$

After simplication,

$$P(T_1=t|Y=y) = \binom{y}{t} \left(\frac{u_1}{u}\right)^t \left(1 - \frac{u_1}{u}\right)^{y-t},$$

which is a Binomial distribution with parameters $(y, u_1/u)$. Since u_{1c} is the mean of the above binomial distribution,

$$u_{1c} = c * \frac{u_1}{u} \quad .$$

Remark : When $c \geq K+1$, u_{1c} can be estimated. If the T's, when restricted to R, are independent, then the hypotheses of Propositions 4.3.1 and 4.3.2 hold.

4.4 Estimation of Parameters in Model 3

The parameters to be estimated are $\underline{u}$ and $\underline{S}_1$. With the same notations as in section 4.2, the estimation process consists of first estimating u (mean of Y, E[Y]) and S^2 (variance of Y, $E[Y-u]^2$) and then expressing u_i and $\underline{S}_1(i,j)$ in terms of these quantities.

The following additional notations are used.
Let h(y) and N(t) be respectively the probability density function and the moment generating function of Y; let g(y) and M(t) be respectively the probability density function and the moment generating function of the conditional random variable $Y|Y\geq c$.

By [10],

$$E(Y|Y\geq c) = M'(0),$$
$$E(Y^2|Y\geq c) = M''(0),$$
$$E(Y^3|Y\geq c) = M'''(0)$$
$$\text{and } N(t) = \exp(ut+S^2t^2/2) \qquad (4.4.1)$$

We now relate u and S to the estimatable quantities

M'(0), M''(0) and M'''(0), making use of the following identity.

$$\frac{d}{dt}\int_{b(t)}^{a(t)} h(y)\,dy = a'(t)h[a(t)] - b'(t)h[b(t)] \quad (4.4.2)$$

Proposition 4.4.1

If Y is a normal random variable with mean u and variance S^2, then

$$u = [\,M'''(0) - M'(0) * (2M''(0)-2cM'(0)+c^2)\,] \,/\, [\,M''(0) - 2(M'(0))^2 + 2cM'(0) - c^2\,]$$

and $$S^2 = M''(0) + u * (c-M'(0)) - cM'(0),$$

where M(t) is the moment generating function of $Y|Y\geq c$.

Proof

$$g(y) = \frac{h(y)}{P(Y\geq c)} = \begin{cases} 0 & \text{if } y < c \\ \dfrac{\frac{1}{(2*PI)^{1/2}S}\exp(\frac{-(y-u)^2}{2S^2})}{P(Y\geq c)} & \text{if } y \geq c. \end{cases}$$

Thus,

$$M(t) = \int_{-\infty}^{\infty} e^{ty} g(y)\,dy$$

$$= \frac{1}{P(Y\geq c)} \frac{\exp(ut+S^2t^2/2)}{(2*PI)^{1/2}S} \int_c^{\infty} \exp(\frac{-(y-u-S^2t)^2}{2S^2})\,dy$$

$$= \frac{N(t)}{P(Y \geq c)} \int_{c-s^2t}^{\infty} \frac{1}{(2*PI)^{1/2}s} \exp\left(\frac{-(y-u)^2}{2s^2}\right) dy \quad \text{by (4.4.1)}$$

$$= N(t)G(t)/P(Y \geq c)$$

where $G(t) = \int_{c-s^2t}^{\infty} \frac{1}{(2*PI)^{1/2}s} \exp\left(\frac{-(y-u)^2}{2s^2}\right) dy$

Differentiating M(t) with respect to t, making use of identity (4.4.2) and setting t=0, the following three equations are obtained.

$$M'(0) = u + \frac{G'(0)}{P(Y \geq c)}$$

$$M''(0) = u^2 + s^2 + (c+u)\frac{G'(0)}{P(Y \geq c)}$$

$$M'''(0) = u^3 + 3us^2 + (2s^2 + uc + u^2 + c^2)\frac{G'(0)}{P(Y \geq c)}$$

$\frac{G'(0)}{P(Y \geq c)}$, u and s^2 can be considered as unknowns in the above equations.

Eliminating $\frac{G'(0)}{P(Y \geq c)}$ from the second and third equations,

$$M''(0) = u^2 + s^2 + (c + u)(M'(0) - u) \qquad (4.4.3)$$

$$M'''(0) = u^3 + 3us^2 + (2s^2 + uc + u^2 + c^2)(M'(0) - u) \qquad (4.4.4)$$

Identity (4.4.3) gives

$$s^2 = M''(0) - u^2 - (c + u)(M'(0) - u)$$

$$= M''(0) + u[c - M'(0)] - cM'(0) \qquad (4.4.5)$$

After substituting (4.4.5) into (4.4.4) and solving for u, the desired result follows.

The above proposition can be used to estimate u and S^2 by setting c = K+1, K+2 etc.

Next, estimate u_i and $\underline{S}_1(i,j)$.

By [11], (T_i, Y), $1 \leq i \leq n$, is bivariate normal, and $(T_i|Y=c)$ is also normal with

$$\text{Mean} = u_i + g_{iY}\frac{S_i}{S}(c-u) \text{ and}$$

$$\text{variance} = S_i^2(1-g_{iY}^2)$$

where $S_i^2 = \text{Var}(T_i)$

g_{iY} = correlation coefficient of T_i and Y.

By definition,

$$u_{ic} = E(T_i|Y=c)$$

$$= u_i + g_{iY}\frac{S_i}{S}(c-u). \qquad (4.4.6)$$

Let a and b be any two values of c, c > K.

Substituting these values in (4.4.6), we obtain

$$g_{iY}\frac{S_i}{S} = \frac{u_{ia}-u_{ib}}{a-b} \qquad (4.4.7)$$

Substituting (4.4.7) into (4.4.6), and solve for u_i,

$$u_i = u_{ic} - \frac{u_{ia}-u_{ib}}{a-b}(c-u) \qquad (4.4.8)$$

(4.4.7) can also be written as

$$g_{iY}S_i = S\frac{u_{ia}-u_{ib}}{a-b} \qquad (4.4.9)$$

Hence (4.4.8) and (4.4.9) permit the estimation of u_i and $g_{iY}S_i$, $1 \leq i \leq n$.

Now,

$$E(T_i^2|Y=c) = Var(T_i|Y=c) - [E(T_i|Y=c)]^2$$
$$= S_i^2(1-g_{iY}^2) - u_{ic}^2$$
$$= S_i^2 - (S_i g_{iY})^2 - u_{ic}^2$$

$$\Rightarrow S_i^2 = E(T_i^2|Y=c) + u_{ic}^2 + \left(S\,\frac{u_{ia}-u_{ib}}{a-b}\right)^2 .$$

Hence S_i^2, $1 \le i \le n$, can also be estimated by setting $c > K$ in the above equation.

The covariance of T_i, T_j and $\underline{S}_1(i,j)$ can be estimated as follows.

Let $S_{ij} = \underline{S}_1(i,j)$

By [11] $(T_i+T_j|Y=c)$ is normal with

$$\text{mean} = (u_i+u_j) + g_{(i+j)Y}\,\frac{S_{(i+j)}}{S}(c-u) \quad \text{and}$$

$$\text{variance} = S_{(i+j)}^2 - [S_{(i+j)} * g_{(i+j)Y}]^2,$$

where $S_{(i+j)}$ is the standard deviation of (T_i+T_j) and $g_{(i+j)Y}$ is the correlation coefficient of (T_i+T_j) and Y.

Applying the same technique as in the estimation of S_i^2,

$$S_{(i+j)}^2 = E\,[(T_i+T_j)^2|Y=c\,] + u_{(i+j)c}^2$$
$$+ \left[\,S * \frac{u_{(i+j)a} - u_{(i+j)b}}{a-b}\right]^2 .$$

where $u_{(i+j)c} = E(T_i+T_j|Y=c)$.

Hence $S_{(i+j)}^2$ can be estimated by setting c = K+1, K+2, etc in the above equation.

But $S_{(i+j)}^2 = Var(T_i+T_j)$

$= Var(T_i) + Var(T_j) + 2Cov(T_i, T_j)$

$= S_i^2 + S_j^2 + 2S_{ij}$

therefore,

$S_{ij} = [S_{(i+j)}^2 - S_i^2 - S_j^2] / 2$ can be estimated.

4.5 Estimation of the Number of Relevance Documents

Let u_i and v_i be the estimated expected weight of the i^{th} term in the relevant documents and the irrelevant documents respectively. Let F_i be the sum of the weights of the i^{th} term in the set of all documents. Then |R| can be estimated by solving the following two equations with unknowns |R| and |I|.

$$u_i * |R| + v_i * |I| = F_i$$

$$|R| + |I| = |D|.$$

Chapter 5

Measurement of the Importance of Index Term with respect to Retrieval Performance in Model 1

5.1 Performance Measure

It is common to have a few hundred terms in a feedback query. Retrieval using so many terms is clearly not economical. Therefore, it is desirable to rank the terms according to their usefulness with respect to the user so that the less useful terms can be discarded. The analysis of ranking terms turns out to be rather involved. Thus analysis is restricted here to the study of the effect of deleting a term in Model 1.

The optimal query Q* in Model 1 is

$$Q^* = (w_1, w_2, \ldots, w_n),$$

$$\text{where } w_i = \log \frac{p_i/1-p_i}{r_i/1-r_i}, \qquad 1 \leq i \leq n.$$

Without loss of generality, assume

$$|w_1| \geq |w_2| \geq \ldots \geq |w_n| \geq 0,$$

and the set of document vectors is

$$D = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_{2^n}\},$$

where each $\underline{x}_i$ has a distinct combination of n terms.

The effect on retrieval performance of deleting term j verses deleting term i will be studied. It will be shown that the i^{th} term is "more" important than the j^{th} term if $|w_i| > |w_j|$. Thus the n^{th} term is the "least important" term in Q*.

Let Q*-{i} be the query identical to Q* except that its i^{th} component is zero, i.e. the i^{th} term of Q* is deleted. Let $i \in \underline{x}$ and $i \notin \underline{x}$ be the events that the i^{th} component of $\underline{x}$ is 1 and 0 respectively.

Retrieval performance is best measured in terms of recall and precision (see chapter 2). An example is presented in section 5.2 which illustrates that Q*-{i} is not better than Q*{j} at all recall levels for any i and j. Thus, there is no "best" query in $\{ Q-\{i\} \mid 1 \leq i \leq n \}$ in terms of recall and precision. However, it will be shown in section 5.3 that Q*-{j} is "better" than Q*-{i} with respect to <u>document ranking</u> if $|w_i| > |w_j|$.

In order to compare the performance of Q*-{i} and Q*-{j}, we first define associates and a measure of goodness of a document. By making use of this measure, the documents retrieved by Q*-{i} will be compared to those of Q*-{j}.

<u>Definition 5.1.1</u>

Document $\underline{x}_m$ is said to be <u>better</u> than document $\underline{x}_p$ if $f(Q^*, \underline{x}_m) > f(Q^*, \underline{x}_p)$.

Let $g(\underline{x})=f(Q^*,\underline{x})$. The value of $g(\underline{x})$ is a measure of "goodness" of document $\underline{x}$. The notion of goodness is consistent with the probability of relevance, $\phi(\underline{x})$, of $\underline{x}$ in chapter 2 and

$$g(\underline{x}_m) > g(\underline{x}_p) \iff \phi(\underline{x}_m) > \phi(\underline{x}_p).$$

Definition 5.1.2

The associate of $\underline{x}_m$ with respect to term i is the vector $\underline{x}_p$ which is identical to $\underline{x}_m$ except at term i. $\{\underline{x}_m, \underline{x}_p\}$ is called a pair of associates with respect to term i and there are 2^{n-1} distinct pairs of associates of term i.

Clearly when $\{\underline{x}_m, \underline{x}_p\}$ are associates with respect to term i, the following properties hold.

(a) $| f(Q^*, \underline{x}_m) - f(Q^*, \underline{x}_p) | = |w_i|$,

(b) $f(Q^*-\{i\}, \underline{x}_m) = f(Q^*-\{i\}, \underline{x}_p)$

$$= \min \{ f(Q^*, \underline{x}_m), f(Q^*, \underline{x}_p) \}$$

i.e. $Q^*-\{i\}$ gives the same rank to $\underline{x}_m$ and $\underline{x}_p$,

(c) if $\{\underline{x}_m', \underline{x}_p'\}$ is another pair of associates with respect to term i, $g(\underline{x}_m) > g(\underline{x}_p)$ and $g(\underline{x}_m') > g(\underline{x}_p')$

then

$$g(\underline{x}_m) > g(\underline{x}_m') \iff g(\underline{x}_p) > g(\underline{x}_p').$$

Definition 5.1.3

If Q_1 retrieves $\{\underline{x}_{1,1}, \underline{x}_{1,2}, \ldots, \underline{x}_{1,k}\}$, a set of k documents,

Q_2 retrieves $\{\underline{x}_{2,1}, \underline{x}_{2,2}, \ldots, \underline{x}_{2,k}\}$, a set of k

documents,

and $g(\underline{x}_{1,m}) \geq g(\underline{x}_{1,m+1})$

$g(\underline{x}_{2,m}) \geq g(\underline{x}_{2,m+1})$, $\quad 1 \leq m < k$

then Q_1 is said to be <u>better than</u> Q_2 <u>with respect to document ranking</u> (written as $Q_1 > Q_2$) if $g(\underline{x}_{1,m}) \geq g(\underline{x}_{2,m})$ for $1 \leq m \leq k$ and at least one inequality is '>' for some k.

5.2 Q*-{i} is not better than Q*-{i} at all recall levels

An optimal retrieval rule retrieves documents in descending order of relevance or "goodness". The deletion of term i from Q* causes the associates with respect to term i to be retrieved together.

Suppose document $\underline{x}$ which does not contain term i is between the two associates $\{\underline{x}_m, \underline{x}_p\}$ with respect to term i, $g(\underline{x}_m) > g(\underline{x}) > g(\underline{x}_p)$. Then deleting term i has no effect on the "retrieval status value" of document $\underline{x}$ but will lower the retrieval status value of $\underline{x}_m$ to that of $\underline{x}_p$. If $w_i > 0$, then $\underline{x}_m$, which has a higher probability of relevance than $\underline{x}_p$, is retrieved after $\underline{x}$. If $w_i < 0$, then $\underline{x}_p$, which has a lower probability of relevance, is retrieved ahead of $\underline{x}$. In either case a loss of precision results. Similarly, it can also be seen that worse retrieval performance will result when document $\underline{x}$ contains term i. Thus the more documents there are between associates of a term, the higher the likelihood of losing precision when the term is deleted. As a consequence, when considering the choice of deleting one

term versus another term, it is desirable to choose the term with as few documents between its associates as possible.

Intuitively, it seems that $Q^*-\{j\}$ would be better than $Q^*-\{i\}$ if $|w_i| > |w_j|$. However the following example illustrates that the precision of $Q^*-\{j\}$ can be lower than that of $Q^*-\{i\}$ at some recall points, though $Q^*-\{j\}$ is actually better with respect to document ranking, as is shown in section 5.3.

Let $Q^* = (w_1, w_2)$

Let $|R(\underline{x})|$ and $|I(\underline{x})|$ be the expected number of relevant and irrelevant documents with representation $\underline{x}$ respectively.

If $\underline{x} = (1,0)$ then $|R(\underline{x})| = |R|*p_1*(1-p_2)$ and

$$|I(\underline{x})| = |I|*r_1*(1-r_2).$$

The performance of $Q^*-\{1\}$ and $Q^*-\{2\}$ will be compared.

Assume $|R| = 100$, $|I| = 1000$ and

	T_1	T_2
p_i	0.90	0.20
r_i	0.75	0.40
w_i	0.48	-0.44

	$f(Q^*,\underline{x})$	$\lvert I(\underline{x})\rvert$	$\lvert R(\underline{x})\rvert$	cum.$\lvert I(\underline{x})\rvert$ / cum.$\lvert R(\underline{x})\rvert$
$\underline{x}_1=(1,0)$	0.48	450	72	450/72 = 6.25
$\underline{x}_2=(1,1)$	0.04	300	18	750/90 = 8.25
$\underline{x}_3=(0,0)$	0.00	150	08	900/98 = 9.18
$\underline{x}_4=(0,1)$	-0.44	100	02	1000/100 = 10.

Associates of term 2 are $\{\underline{x}_1, \underline{x}_2\}$ and $\{\underline{x}_3, \underline{x}_4\}$

Associates of term 1 are $\{\underline{x}_1, \underline{x}_3\}$ and $\{\underline{x}_2, \underline{x}_4\}$

Q^*-{2} ranks $\{\underline{x}_1, \underline{x}_2\}$ highest.

Q^*-{1} ranks $\{\underline{x}_1, \underline{x}_3\}$ highest.

and $g(\underline{x}_2) > g(\underline{x}_3)$.

The number of relevant documents with representation $\underline{x}_1$ and $\underline{x}_2$ is 90.

The number of irrelevant documents with representation $\underline{x}_1$ and $\underline{x}_2$ is 750.

The number of relevant documents with representation $\underline{x}_1$ and $\underline{x}_3$ is 80.

The number of irrelevant documents with representation $\underline{x}_1$ and $\underline{x}_3$ is 600.

If only 80 relevant documents are recalled, Q^*-{1} retrieves 600 irrelevant ones, but Q^*-{2} retrieves 750*80/90=666 irrelevant ones.

Q^*-{2} is not better than Q^*-{1} at this recall level. The reason is that a random sample has to be drawn from the set $\{\underline{x}_1, \underline{x}_2\}$, since all documents in the set has the same retrieval status value with respect to Q^*-{2}.

5.3 Which is better, Q*-{j} or Q*-{i} ?

In this section Q*-{j} is shown to be better than Q*-{i} with respect to document ranking (Definition 5.1.3) if $|w_i| > |w_j|$. As a result, Q*-{n} ranks documents closer to the optimal ranking by Q* than does Q*-{i}, $1 \leq i < n$.

First, a series of lemmas about the properties of the best document with respect to Q* are proved. Second, Proposition 5.3.4 relates the documents retrieved by Q*-{i} and Q*. Third, Proposition 5.3.6 illustrates the relationship between the associates of term i and the associates of term j, making use of Lemma 5.3.5. Lastly, the main result, Proposition 5.3.7, shows that Q*-{j} is better with respect to document ranking.

We further assume no two distinct document vectors have the same retrieval status value with respect to Q*. Without loss of generality, assume

$$g(\underline{x}_1) > g(\underline{x}_2) > \ldots > g(\underline{x}_n).$$

The above assumption guarantees, for any i, $1 \leq i \leq n$, that $f(Q^*-\{i\}, \underline{x}_m) = f(Q^*-\{i\}, \underline{x}_p) \Rightarrow \underline{x}_m, \underline{x}_p$ are associates with respect to term i.

The following lemma shows that the best document $\underline{x}_1$ is retrieved ahead of any other document by Q*-{i}, $1 \leq i \leq n$.

Lemma 5.3.1

For any i, $1 \leq i \leq n$,

$f(Q^*-\{i\}, \underline{x}_1) \geq f(Q^*-\{i\}, \underline{x}_t)$ for $1 \leq t \leq 2^n$.

Proof

Consider the case when $w_i > 0$ (the case where $w_i < 0$ can be handled in a similar way).

Let $\underline{x}_v$ be an associate of $\underline{x}_1$ with respect to term i. Then $i \in \underline{x}_1$, and $i \notin \underline{x}_v$. Suppose there is a document $\underline{x}_m$ lying between $\underline{x}_1$ and $\underline{x}_v$,

i.e. $g(\underline{x}_1) > g(\underline{x}_m) > g(\underline{x}_v)$.

$\underline{x}_m$ must contain term i, otherwise its associate with respect to term i would be better than $\underline{x}_1$.

Now $f(Q^*, \underline{x}_1) = f(Q^*-\{i\}, \underline{x}_1) + w_i$

$f(Q^*, \underline{x}_m) = f(Q^*-\{i\}, \underline{x}_m) + w_i$

$g(\underline{x}_1) > g(\underline{x}_m) \Rightarrow f(Q^*-\{i\}, \underline{x}_1) > f(Q^*-\{i\}, \underline{x}_m)$.

Hence any document better than $\underline{x}_v$ satisfies the inequality. Next consider any document $\underline{x}_t$ that is worse than $\underline{x}_v$.

$$\begin{aligned} f(Q^*-\{i\}, \underline{x}_1) &= f(Q^*-\{i\}, \underline{x}_v) \\ &= f(Q^*, \underline{x}_v) \\ &> f(Q^*, \underline{x}_t) \\ &\geq f(Q^*-\{i\}, \underline{x}_t). \end{aligned}$$

Lemma 5.3.2

The associate of $\underline{x}_1$ with respect to term n must be $\underline{x}_2$ if $|w_i| > |w_n|$, $1 \leq i < n$.

Proof

For simplicity, assume $w_n > 0$.

In order to be the document of highest relevance, $\underline{x}_1$ must contain all the positive terms and none of the negative terms. Any $\underline{x}$ different from $\underline{x}_1$ must contain at least one negative term or be missing at least one positive term. In either case, $f(Q^*, \underline{x}_1) - f(Q^*, \underline{x}) \geq w_n$ and the result follows.

As a consequence of the above two lemmas, the first pair of documents retrieved by $Q^*-\{n\}$ is $\{\underline{x}_1, \underline{x}_2\}$. This indicates that $Q^*-\{n\}$ is better than $Q^*-\{i\}$, $1 \leq i \leq n-1$, with respect to document ranking when retrieving the first pair of documents. The same result is true when comparing $Q^*-\{j\}$ with $Q^*-\{i\}$, for $|w_i| > |w_j|$. This is demonstrated by Lemma 5.3.3, whose proof is trivial.

Lemma 5.3.3

Let $\underline{x}_{i,2}$ and $\underline{x}_{j,2}$ be the associates of $\underline{x}_1$ with respect to term i and j respectively, then

$$g(\underline{x}_{j,2}) > g(\underline{x}_{i,2}) \Leftrightarrow |w_i| > |w_j|.$$

The following proposition relates the documents retrieved by $Q^*-\{i\}$ and Q^*. As a consequence of the proposition, it can be shown easily by induction on k that whenever $Q^*-\{i\}$ retrieves 2k documents, the 2k documents must include $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_k\}$.

Proposition 5.3.4

Let $\{\underline{x}_{i,1}, \underline{x}_{i,2}, \cdots, \underline{x}_{i,2k}\}$ be the set of 2k documents having the highest retrieval status value with respect to $Q^*-\{i\}$, for some i, $1 \leq i \leq n$.

If $\{\underline{x}_m, \underline{x}_p\}$ is a pair of associates of term i, $g(\underline{x}_m) > g(\underline{x}_p)$ and $f(Q^*-\{i\}, \underline{x}_{i,t}) \geq f(Q^*-\{i\}, \underline{x}_p)$ $\quad 1 \leq t \leq 2k$,

then

(i) $g(\underline{x}_{i,t}) \geq g(\underline{x}_p), \quad 1 \leq t \leq 2k,$

(ii) $\{\underline{x}_{i,1}, \underline{x}_{i,2}, \cdots, \underline{x}_{i,2k}\}$
$= \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_m,$ (2k-m-1) documents between $\underline{x}_m$ and $\underline{x}_p$, $\underline{x}_p\}$

i.e. $g(\underline{x}) > g(\underline{x}_m)$
$\Rightarrow f(Q^*-\{i\}, \underline{x}) > f(Q^*-\{i\}, \underline{x}_m)$.

Proof

(i)

$$g(\underline{x}_{i,t}) = f(Q^*, \underline{x}_{i,t})$$
$$= f(Q^*-\{i\}, \underline{x}_{i,t}) + d_1 w_i$$
$d_1=1$ if $i \in \underline{x}_{i,t}$, $d_1=0$ if $i \notin \underline{x}_{i,t}$
$$\geq f(Q^*-\{i\}, \underline{x}_p) + d_1 w_i$$
$$= f(Q^*, \underline{x}_p) - d_2 w_i + d_1 w_i$$
$d_2=1$ if $w_i<0$, $d_2=0$ if $w_i>0$
$$= f(Q^*, \underline{x}_p) + (d_1-d_2) w_i$$
$$\geq f(Q^*, \underline{x}_p).$$

(ii)

$g(\underline{x}) > g(\underline{x}_m)$

$\Leftrightarrow f(Q^*,\underline{x}) > f(Q^*, \underline{x}_m)$

$\Leftrightarrow f(Q^*-\{i\}, \underline{x}) + d_1 w_i = f(Q^*, \underline{x}) > f(Q^*, \underline{x}_m)$

$d_1 = 1$ if $i \in \underline{x}$, $d_1 = 0$ if $i \notin \underline{x}$.

$\Rightarrow f(Q^*-\{i\}, \underline{x}) + d_2 w_i \geq f(Q^*, \underline{x}) > f(Q^*, \underline{x}_m)$

$d_2 = 0$ if $w_i < 0$, $d_2 = 1$ if $w_i > 0$

$\Leftrightarrow f(Q^*-\{i\}, \underline{x}) + d_2 w_i > f(Q^*, \underline{x}_m) = f(Q^*-\{i\}, \underline{x}_m) + d_2 w_i$

$\Leftrightarrow f(Q^*-\{i\}, \underline{x}) > f(Q^*-\{i\}, \underline{x}_m)$.

The following lemma can be proved easily by induction on n. It is used to prove Proposition 5.3.6.

Lemma 5.3.5

Let $S_1 = \{a_1, a_2, \ldots, a_n\}$ and $S_2 = \{b_1, b_2, \ldots, b_n\}$ be two sets of real numbers of arbitrary size n.
If $a_k \geq b_k$ then $a_{(k)} \geq b_{(k)}$ for $1 \leq k \leq n$
where $a_{(k)}$ and $b_{(k)}$ are the k^{th} largest numbers in S_1 and S_2 respectively.

The next proposition compares the k^{th} "best" associate pair of term i with the k^{th} "best" associate pair of term j. There are 2^{n-1} associate pairs of term i and 2^{n-1} associate pairs of term j. We can compare them pairwisely by first rearrange the associate pairs in their order of retrieval status value w.r.t. $Q^*-\{i\}$ and $Q^*-\{j\}$ respectively.

Proposition 5.3.6

Let $\{ \{\underline{x}_{i,2k-1}, \underline{x}_{i,2k}\} : 1 \leq k \leq 2^{n-1} \}$ and

$\{ \{\underline{x}_{j,2k-1}, \underline{x}_{j,2k}\} : 1 \leq k \leq 2^{n-1} \}$

be the 2^{n-1} associate pairs of term i and term j respectively.

If the associate pairs are rearranged such that

$$g(\underline{x}_{i,2k-1}) > g(\underline{x}_{i,2k}),$$

$$g(\underline{x}_{j,2k-1}) > g(\underline{x}_{j,2k}), \quad 1 \leq k \leq 2^{n-1},$$

$$f(Q^*-\{i\}, \underline{x}_{i,2k-1}) > f(Q^*-\{i\}, \underline{x}_{i,2k+1}),$$

$$f(Q^*-\{j\}, \underline{x}_{j,2k-1}) > f(Q^*-\{j\}, \underline{x}_{j,2k+1}), \quad 1 \leq k \leq 2^{n-1}-2$$

and $|w_i| > |w_j|$

then

$$g(\underline{x}_{i,2k-1}) \geq g(\underline{x}_{j,2k-1}),$$

and $g(\underline{x}_{i,2k}) \leq g(\underline{x}_{j,2k}), \quad 1 \leq k \leq 2^{n-1}.$

The propostition states that the better associate, $\underline{x}_{i,2k-1}$, with respect to term i is better, than the better associate, $\underline{x}_{j,2k-1}$, with respect to term j but the worse associate, $\underline{x}_{i,2k}$, with respect to term i is worse than the worse associate, $\underline{x}_{j,2k}$, with respect to term j. Thus, it is still not clear whether $Q^*-\{j\}$ is better than $Q^*-\{i\}$.

Sketch of proof

(a) construct a simple one-to-one mapping M from D to D such that if $\{\underline{x}_m, \underline{x}_p\}$ are associates with respect to term j and $g(\underline{x}_m) > g(\underline{x}_p)$

then (i) $\{M(\underline{x}_m), M(\underline{x}_p)\}$ are associates w.r.t. term i

(ii) $g(M(\underline{x}_m)) > g(M(\underline{x}_p))$

(iii) $g(\underline{x}_m) \leq g(M(\underline{x}_m))$

(iv) $g(\underline{x}_p) \geq g(M(\underline{x}_p))$

(b) set

$$a_t = g(\underline{x}_m)$$
$$b_t = g(M(\underline{x}_m))$$
$$c_t = g(\underline{x}_p)$$
$$d_t = g(M(\underline{x}_p))$$

Then $a_t \leq b_t$, $c_t \geq d_t$; $a_t > c_t$, $b_t > d_t$ by definition of the mapping M.

(c) By Lemma 5.3.5,

$$a_{(k)} \leq b_{(k)}$$
$$c_{(k)} \geq d_{(k)} \qquad 1 \leq k \leq 2^{n-1}$$

where $a_{(k)}$ is the k^{th} largest number among the a_t's.

Since $a_t - c_t = |w_j|$ and $b_t - d_t = |w_i|$ for all t, we have

$a_{(k)} = a_t$ for some t $\Rightarrow$ $c_{(k)} = c_t$;

$b_{(k)} = b_t$ for some t $\Rightarrow$ $d_{(k)} = d_t$.

Hence $a_k > c_k \Leftrightarrow a_{(k)} > c_{(k)}$

$b_k > d_k \Leftrightarrow b_{(k)} > d_{(k)}$, $1 \leq k \leq 2^{n-1}$.

(d) show that the sorting of a's, b's, c's and d's result in an arrangement of the associate pairs, more precisely,

$$g(\underline{x}_{i,2k-1}) = b_{(k)} > d_{(k)} = g(\underline{x}_{i,2k})$$
$$g(\underline{x}_{j,2k-1}) = a_{(k)} > c_{(k)} = g(\underline{x}_{j,2k})$$

$$b_{(k)} > b_{(k+1)} \iff f(Q^*-\{i\}, \underline{x}_{i,2k-1}) > f(Q^*-\{i\}, \underline{x}_{i,2k+1})$$
$$a_{(k)} > a_{(k+1)} \iff f(Q^*-\{j\}, \underline{x}_{j,2k-1}) > f(Q^*-\{j\}, \underline{x}_{j,2k+1})$$

(e) finally show that the results are true,

$$a_{(k)} \leq b_{(k)} \iff g(\underline{x}_{j,2k-1}) \leq g(\underline{x}_{i,2k-1})$$
$$c_{(k)} \geq d_{(k)} \iff g(\underline{x}_{j,2k}) \geq g(\underline{x}_{i,2k})$$

Proof :

There are 4 cases :

(i) $w_i > w_j > 0$

(ii) $w_i < w_j < 0$

(iii) $w_i > 0$, $w_j < 0$, $w_i+w_j > 0$

(iv) $w_i < 0$, $w_j > 0$, $w_i+w_j < 0$

A mapping M for cases (i) and (ii) is

$M : D \dashrightarrow D$

where $M(\underline{x})$ is the vector formed by interchanging the i^{th} and j^{th} component of $\underline{x}$.

i.e. $\underline{x} = (\ldots, x_i, \ldots, x_j, \ldots)$

$M(\underline{x}) = (\ldots, x_j, \ldots, x_i, \ldots)$

This mapping M can be shown to satisfy condition (a).

A mapping M for case (iii) and (iv) is

$M : D \dashrightarrow D$

where $M(\underline{x})$ is the vector formed by interchanging the i^{th} and j^{th} component of $\underline{x}$ and then complementing them.

i.e. $\underline{x} = (\ldots, x_i, \ldots, x_j, \ldots)$

$M(\underline{x}) = (\ldots, 1-x_j, \ldots, 1-x_i, \ldots)$

This mapping M can also be shown to satisfy condition (a).

As an illustration, the proof of case (iv) is as follows :

(iv) $w_i < 0$, $w_j > 0$, $w_i+w_j < 0$

$\underline{x} = (\ldots, e_i, \ldots, e_j, \ldots)$

$M(\underline{x}) = (\ldots, (1-e_j), \ldots, (1-e_i), \ldots)$

Let $\{\underline{x}_m, \underline{x}_p\}$ be a pair of associates of term j,

such that $g(\underline{x}_m) > g(\underline{x}_p)$ $(\Rightarrow j \in \underline{x}_m, j \notin \underline{x}_p)$.

Clearly, $\{M(\underline{x}_m), M(\underline{x}_p)\}$ is a pair of associates of term i.

$$g(M(\underline{x}_m)) - g(M(\underline{x}_p)) = w_i(1-1) - w_i(1-0)$$
$$= -w_i > 0.$$

$$g(\underline{x}) - g(M(\underline{x})) = w_i e_i + w_j e_j - w_i(1-e_j) - w_j(1-e_i)$$

$$g(\underline{x}_m) - g(M(\underline{x}_m)) = w_i e_i + w_j * 1 - w_i(1-1) - w_j(1-e_i)$$
$$= (w_i+w_j) * e_i \leq 0$$

$$g(\underline{x}_p) - g(M(\underline{x}_p)) = w_i e_i + w_j * 0 - w_i(1-0) - w_j(1-e_i)$$
$$= (w_j+w_i) * (e_i-1) \geq 0$$

Conditions of (a) are satisfied.

Step (b) to step (e) are straight forward and the results immediately follow.

The last proposition shows that $Q^*-\{j\} > Q^*-\{i\}$ with respect to document ranking if $|w_i| > |w_j|$.

Proposition 5.3.7

If the set of 2k documents retrieved by $Q^*-\{j\}$ and $Q^*-\{i\}$ are

$\{ \underline{x}_{j,1}, \underline{x}_{j,2}, \cdots, \underline{x}_{j,2k} \}$ and

$\{ \underline{x}_{i,1}, \underline{x}_{i,2}, \cdots, \underline{x}_{i,2k} \}$ respectively,

and $g(\underline{x}_{j,h}) > g(\underline{x}_{j,h+1})$,

$g(\underline{x}_{i,h}) > g(\underline{x}_{i,h+1})$, $1 \leq h < 2k$

then

$g(\underline{x}_{j,t}) \geq g(\underline{x}_{i,t})$, $1 \leq t \leq 2k$.

Proof

(By induction on k)

At k=1.

By Lemma 5.3.1 and Lemma 5.3.3, the proposition is true at k=1.

Suppose the proposition is true for 1 to k.

Let the set of 2k documents retrieved by Q*-{j} and Q*-{i} be $\{\underline{x}_{j,1}, \underline{x}_{j,2}, \cdots, \underline{x}_{j,2k}\}$ and $\{\underline{x}_{i,1}, \underline{x}_{i,2}, \cdots, \underline{x}_{i,2k}\}$ respectively.

Consider the case for k+1.

Let the next pair of documents retrieved by Q*-{j} and Q*-{i} be $\{\underline{x}_m, \underline{x}_p\}$ and $\{\underline{x}_t, \underline{x}_u\}$ respectively, such that $g(\underline{x}_m) > g(\underline{x}_p)$ and $g(\underline{x}_t) > g(\underline{x}_u)$.

By Proposition 5.3.4, $g(\underline{x}_{j,2k}) > g(\underline{x}_p)$ and $g(\underline{x}_{i,2k}) > g(\underline{x}_u)$.

By Proposition 5.3.6, $g(\underline{x}_p) \geq g(\underline{x}_u)$ but $g(\underline{x}_m) \leq g(\underline{x}_t)$.

By Proposition 5.3.4, the set of 2k documents retrieved by Q*-{i} must be

$\{\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_{t-1}$, 2k-t documents between $\underline{x}_t$ and $\underline{x}_{i,2k}$, $\underline{x}_{i,2k}\}$,

and by induction assumption and Proposition 5.3.4, the set of 2k documents retrieved by Q*-{j} must be

$\{\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_{t-1}, \cdots, \underline{x}_{m-1}$, 2k-m documents between $\underline{x}_m$ and $\underline{x}_{j,2k}, \underline{x}_{j,2k}\}$.

After the inclusion of $\{\underline{x}_m, \underline{x}_p\}$ and $\{\underline{x}_t, \underline{x}_u\}$ respectively, the ordering is clearly preserved.

Corollory 5.3.7

Q*-{n} ranks documents closer to optimal than Q*-{i}, $1 \leq i < n$.

Hence we can rank the usefulness of terms in their order of $|w_i|$.

Chapter 6

Conclusion

The thesis identifies the objective of retrieving more documents that are of interest to the user to be a two-way classification or discriminant problem. Thus an optimal retrieval rule is one that best discriminates the set of relevant documents from the set of irrelevant documents. Two of the most common components in retrieval performances are precision and recall. They are used to evaluate the performance of retrieval rules and queries in the thesis. An optimal retrieval rule (2.3.1) has been derived that maximizes precision at any recall and has been shown that it also ranks documents in descending order of relevance.

An information retrieval system can be implemented more easily by queries and a simple matching function (2.1.1) rather than the optimal retrieval rule. One may consider an optimal query to be a linear reduction of the optimal retrieval rule. Unfortunately, such a reduction is not possible in general but under assumptions on the distribution of index terms. Thus the retrieval rule reduces to different query forms under different assumptions. Three common statistical models of information retrieval systems have been studied in the thesis. In

chapter 3, the optimal queries of these three models are derived. Model 1, the binary independent model, has been shown to be a linear approximation of the binary model that incorporates the dependence of terms (Appendix 1).

Constructions of optimal queries are not possible unless there are statistical information of the set of relevant documents and the set of irrelevant documents. We have pointed out that users' initial queries are usually not optimal and the system may be requested to modify the initial queries to achieve better performance. In chapter 4, the processes to estimate the parameters needed to construct the optimal queries of the three models are presented, by making use of relevance information of the documents retrieved by the initial queries. The estimation processes presented here allow limited dependencies of terms but they demonstrate a systematic statistical use of relevance information. It is hoped that the estimation can be generalized to the term dependence case. The optimal query of model 3, which is a special case of the Fisher's linear discriminant, we think, can be applied to non-binary models and more general estimation procedures can be investigated.

Last but not the least, the relative importance of index terms in a query of the first model with respect to retrieval performance has been studied. An example (5.2) has been given to illustrate that deleting the term of

smallest absolute weight from the optimal query is not the best choice in terms of precision and recall when a term has to be deleted. Therefore, a new method - document ranking (5.1.3) is proposed tc compare performance of queries. Section 5.3 shows that the query which results from deleting the term of smallest absolute weight from the optimal query, is best according to this new performance measure (definition 5.1.3), results in least disturbance in document ranking, and ranks the documents closer to the optimal than deleting any other term. Hence the larger the absolute weight of a term the more importance it is in retrieval. Terms can then be ranked in decreasing order of usefulness, allowing the less useful ones to be deleted without seriously affecting retrieval performance. The analysis of ranking terms turns out to be rather involved. It is hoped that the same approach can be generalized to the case of deleting more index terms and to other models. The deletion of terms from feedback queries is necessary in order to speed up the process of retrieval.

There are indeed many questions and problems in the area of feedback query construction but we believe that the theory presented in this thesis demonstrates a systematic statistical study of information retrieval systems and a statistical use of relevance information. Moreover, the study of term deletion is also very important. Though we have made a good start in analysing the usefulness of index terms, further investigation is required.

References

[1] Afifi, A.A. and Azen, S.P. "Statistical Analysis - A Computer Oriented Approach" Academic Press 1972.

[2] Bookstein A. and Kraft, D. "Operational Research Applied to Document Indexing and Retrieval Decisions" JACM 24:418-427, 1977.

[3] Bookstein, A. and Swanson D. "Probabilistic Models of Indexing" J. Amer. Soc. Inform. Sci., 25:312-319, 1974.

[4] Borko, B. and Bernick, M. "Automatic Document Classification, part 2" JACM 11:138-151, 1963.

[5] Duda, R.O. and Hart, P.E. Pattern Classification and Scene Analysis. Wiley 1973.

[6] Fukunage, K. Introduction to Statistical Pattern Recognition. Academic Press, 1972.

[7] Harter,S. "A Probabilistic Approach to Modern Keyword Indexing, part 1" J. Amer. Soc. Inform. Sci. 26:197-206, 1975.

[8] Heine, M.H. "Design Equations for Retrieval Systems Based on the Swets Model" J. Amer. Soc. Inform. Sci. 25:183-198, 1974.

[9] Kraft, D.H. and Bookstein, A. "Evaluation of Information Retrieval Systems : A Decision Theory Approach" J. Amer. Soc. Inform. Sci. 29:31-34, 1978.

[10] Larson H.J. Introduction to Probability Theory and Statistical Inference. 2nd e.d. Wiley, 1974.

[11] Miller, K.S. Multidimensional Gaussian Distribution. John Wiley & Sons, 1964.

[12] Ossorio, P.G. "A Multivariate Procedure for Automatic Document Indexing and Retrieval" Multivariate Behavioral Research 1:479-524, 1966.

[13] Robertson, S.E. and Sparck Jones, K.S. "Relevance Weighting of Search Terms" J. Amer. Soc. Inform. Sci. 27:129-146, 1976.

[14] Rocchio, J.J. and Salton, G. "Information Search Optimization and Iterative Retrieval Techniques" Proc. AFIPS 1965 FJCC, Part 1, 27:293-305.

[15] Salton, G. The SMART Retrieval System - Experiments in Automatic Text Processing. Prentice-Hall, Englewood Cliffs, N.J., 1971.

[16] Swets, J.A. "Information retrieval Systems" Science 141:245-250, 1963.

[17] Van Rijsbergen, C.J. "A Theoretical basis for the use of co-occurrence data in Information Retrieval" Journal of Documentation 33:106-119, 1977.

[18] Yu, C.T., Luk, W.S. and Cheung, T.Y. "A Statistical Model for Relevance Feedback in Information Retrieval" JACM 23:273-286, 1976.

[19] Yu, C.T., Luk, W.S. and Siu, M.K. "On the Estimation of the Number of Desired Records with respect to a Given Query" ACM Transactions on Database Systems, 3:41-44, 1978.

[20] Yu, C.T. and Salton, G. "Precision Weighting - An Effective Automatic Indexing Method" JACM 23:76-78, 1976.

[21] Yu, C.T., Salton, G. and Siu, M.K. "Effective Automatic Indexing Using Term Addition and Deletion" JACM 25:210-225, 1978.

Appendix 1

The optimal query of Model 1 is a "first order" or linear approximation of optimal queries obtainable in a binary model which incorporates the dependence of terms.

The operative part of the optimal retrieval rule, $P(\underline{T}=\underline{x}|C_1)$, can be written in the form of a series expansion [5],

$$P(\underline{T}=\underline{x}|C_1) = P_1(\underline{T}=\underline{x}|C_1) * [\, 1 + \overline{\sum_{i<j}} g_{ij}z_iz_j + \overline{\sum_{i<j<k}} g_{ijk}z_iz_jz_k + \ldots \,]$$

where

$$g_{ij} = E(z_iz_j), \quad g_{ijk} = E(z_iz_jz_k) \ldots$$

$$z_i = \frac{x_i - p_i}{[p_i(1-p_i)]^{1/2}},$$

and $$P_1(\underline{T}=\underline{x}|C_1) = \prod_{i=1}^{n} p_i^{x_i} * (1-p_i)^{1-x_i}$$

i.e. the expansion of $P(\underline{T}=\underline{x}|C_1)$ when T_i's are independent.

When y is small, log(1+y) can be approximated by y, therefore,

$$\log P(\underline{T}=\underline{x}|C_1) = \log P_1(\underline{T}=\underline{x}|C_1) + \overline{\sum_{i<j}} g_{ij}z_iz_j + \ldots$$

and log $P_1(\underline{T}=\underline{x}|C_1)$ is a "first order" or linear approximation of log $P(\underline{T}=\underline{x}|C_1)$.

Thus the optimal query of Model 1 (derived from $\log(P_1(\underline{T}=\underline{x}|C_1)/P_1(\underline{T}=\underline{x}|C_2))$) is a "first order" or linear approximation of the optimal queries obtainable in a binary model which incorporates the dependence of terms.